PB83-232090

Increased Confidence in Spectrum Matching by Use of a Retention Time Library

(U.S.) Environmental Research Lab., Athens, GA

1983

U.S. Department of Commerce
National Technical Information Service
NTIS

TECHNICAL REPORT DATA
(Please read Instructions on the reverse before completing)

1. REPORT NO. EPA-600/D-83-084	2.
3. RECIPIENT'S ACCESSION NO. PB83-232090	
4. TITLE AND SUBTITLE Increased Confidence in Spectrum Matching by Use of a Retention Time Library	5. REPORT DATE 1983 6. PERFORMING ORGANIZATION CODE
7. AUTHOR(S) Walter M. Shackelford	8. PERFORMING ORGANIZATION REPORT NO.
9. PERFORMING ORGANIZATION NAME AND ADDRESS Environmental Research Laboratory U.S. Environmental Protection Agency College Station Road Athens GA 30613	10. PROGRAM ELEMENT NO. ABWD1A 11. CONTRACT/GRANT NO.
12. SPONSORING AGENCY NAME AND ADDRESS Environmental Research Laboratory--Athens GA Office of Research and Development U.S. Environmental Protection Agency Athens GA 30613	13. TYPE OF REPORT AND PERIOD COVERED 14. SPONSORING AGENCY CODE EPA/600/01
15. SUPPLEMENTARY NOTES	

16. ABSTRACT

To successfully extract the maximum amount of information, all dimensions of the gas chromatography/mass spectrometry (GC/MS) data from a sample run must be used. In this work, retention data were combined with reference mass spectra for computer-aided identification of organics in industrial effluent. Use of retention data proved to be a great help in increasing the analyst's confidence in compound identification from low quality spectrum matches. Even greater confidence will be achieved when libraries that include capillary column retention data and gas phase infrared spectra are available.

17. KEY WORDS AND DOCUMENT ANALYSIS		
a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group
18. DISTRIBUTION STATEMENT RELEASE TO PUBLIC	19. SECURITY CLASS *(This Report)* UNCLASSIFIED 20. SECURITY CLASS *(This page)* UNCLASSIFIED	21. NO. OF PAGES 22 22. PRICE

EPA Form 2220-1 (9-73)

EPA-600/D-83-084

PB83-232090

Increased Confidence in Spectrum Matching by Use of a Retention Time Library

by

Walter M. Shackelford

U.S. Environmental Protection Agency

College Station Road

Athens, GA 30613

REPRODUCED BY
NATIONAL TECHNICAL
INFORMATION SERVICE
U.S. DEPARTMENT OF COMMERCE
SPRINGFIELD, VA. 22161

NOTICE

This document has been reviewed in accordance with U.S. Environmental Protection Agency policy and approved for publication. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

ABSTRACT

To successfully extract the maximum amount of information, all dimensions of the gas chromatography/mass spectrometry (GC/MS) data from a sample run must be used. In this work, retention data were combined with reference mass spectra for computer-aided identification of organics in industrial effluent. Use of retention data proved to be a great help in increasing the analyst's confidence in compound identifications from low quality spectrum matches. Even greater confidence will be achieved when libraries that include capillary column retention data and gas phase infrared spectra are available.

Increased Confidence in Spectrum Matching by Use of a Retention Data Library

Introduction. The data acquired in a scanned gas chromatography/mass spectrometry (GC/MS) analysis has three dimensions of qualitative information (Figure 1). Each dimension can provide the chemist with varying degrees of confidence in identification, but it is when these parameters are combined that the power of GC/MS is evidenced. The elution time of a component taken by itself provides useful qualitative information only if the sample has no interferences. Likewise, the masses recorded in a given scan, while providing more qualitative information than elution time, are of little value unless coupled with intensities. The detector's response to a compound without elution time or specific mass data provides little in the way of qualitative information.

Even if two dimensions are combined, the resulting qualitative information falls far short of the total capabilities of GC/MS in qualitative analysis. For instance, an extracted ion current profile (EICP), which combines mass and retention data, while narrowing a chemist's search for compounds having a certain characteristic mass, still requires manual search of each occurrence of that mass for compound identification. Use of the full mass spectrum, which includes mass and intensity for all masses recorded, eliminates many of the ambiguities found in using characteristic ions or retention times alone, but requires that probability based matching (PBM), an automated library search system, be used for acceptable efficiency.

The use of automated spectrum search and retrieval systems is a great aid to qualitative analysis of large numbers of unknowns, but the reliability of such systems is suspect. For example, it is well known that mass spectra, while highly characteristic of a molecule, are not always unique. Also, when dealing with real world data, one often finds contaminated spectra and spectra skewed from instrumental problems.

In this work, elution time information was added to full-scan mass spectra to increase the reliability of automated spectrum matching. In this way, mass spectrum ambiguities were alleviated by requiring retention data matching. Retention data overlaps were overcome by use of mass spectral data. A dynamic historical library was managed that increased in size as compounds were authenticated and more retention data were added. The increase in confidence of automated spectrum matching gained through the use of retention data was measured. In Figure 2 the use of mass, intensity and retention data to narrow the choices for identification is depicted as a set of filters.

Experimental. This work is the result of a study of wastewater from 21 industrial categories and finished water from publicly owned treatment works (POTW). The study encompassed some 4000 samples that were analyzed at 14 contract laboratories and EPA regional laboratories. Details of the study, computer system, and data reduction systems can be found elsewhere[1].

To create the historical library, several homologous series standards were analyzed to provide retention data. The classes of compounds used are shown in Table 1.

The logic of spectrum analysis is shown in Figure 3. The library[2] of reference mass spectra, which contained more than 40,000 spectra, was searched first. In this way, compounds for which no retention data existed in the historical library were not prematurely eliminated from the search procedure.

As compounds are tentatively identified using spectra from different GC columns from different laboratories, provision in the library must be made to differentiate among retention data from different columns. In addition, care must be exercised to differentiate among the internal standards used for reference in retention data.

A search was conducted in three steps:

1. The Chemical Abstracts Service (CAS) number of the best acceptable spectrum match for the unknown was recorded, along with the GC column identification number, relative retention time and internal standard.

2. The historical library (ordered on CAS number) was searched for the candidate's CAS number. If the numbers matched, the GC column and internal standard had to match as well.

3. The relative retention time of the candidate compound and the retention window allowed for the library entry were compared. If these two were in accord a match was recorded.

If the two did not match, the next best acceptable spectrum match was carried through. If no other spectrum matches were acceptable, the computer program flagged the spectrum so that a chemist could make an appropriate decision on its identification.

The Probability Based Matching (PBM) system has been reported by Pesyna and coworkers.[3] Figure 4 depicts the matching parameters used in this study. The measure of match overall quality, K, is theoretically unbounded, but, in practice depends on the number of peaks available for matching in a spectrum. Thus, K_{max}, the K for a perfect match, varies with the number of fragments. This means that rather than relying on K alone, ΔK, the algebraic difference between K_{max} and K, should be considered. To consider K and ΔK simultaneously, the ratio $\Delta K/K_{max}$ can be used as a matching parameter that reflects both positive and negative match qualities.

Confirmation of computer matches was accomplished by reanalysis of sample extracts using capillary column GC/MS[4]. More than 3000 computer-matched identifications were studied in the comfirmation process.

Results and discussion. In Table 2 are shown the relations between the match parameters K and K/K_{max} and the precision of relative retention data. The standard deviation of the relative retention data is expressed in relative retention units and is calculated using data from the 14 participating laboratories. The compounds shown here are representative of commonly found compound classes for each fraction or column. The fused silica capillary

column data are from compounds not commonly found on the acid or base/neutral packed columns.

In comparing the relative retention times with K values for each compound, one can see that whereas retention data variance is very small, the K value range is a factor of 2-5. The narrow retention time windows allow greater confidence in poor spectrum match parameters as will be shown below.

Note that the standard deviations of the fused silica capillary column data are much smaller than those of the other columns. Although a smaller number of laboratories is represented in the fused silica data (only 2 compared to 14 with the packed column data), later work[5] has shown the interlab precision of retention data using fused silica columns to be excellent also.

In Figure 5, the effect that retention data has on matching confidence when using K as the deciding match parameter is seen. The upper curves refer to data collected in this work, whereas the lower curve refers to Atwater's previous study[5] using spectrum data alone. As can be seen, although the curves merge at high K, much higher confidence can be placed in the matches from the present study that have retention time corroboration at low K values.

An anomaly can be seen in the two uppermost curves. Whereas the confirmation rate should increase with increasing K, one of the curves shows a decrease in confirmation rate at high K. Examination of the data revealed that, although aliphatic carboxylic

acids were the largest group of compounds at K>100, only 32 percent were confirmed. Likewise, aromatic acids had a confirmation rate of only 37 percent. Results of the carboxylic acids were deleted from the data set and replotted to obtain the topmost curve, which follows closely that of Atwater[5] at high K.

Poor confirmation rates for the carboxylic acid matches can be attributed to degradation of sample extracts used for confirmation studies during storage. To examine the confirmation characteristics of other compound classes, note Table 3. Even though the ambiguity of matching results among hydrocarbons should be high, hydrocarbons evidence the highest confirmation rates. Perhaps storage stability explains this fact as well. Note that carboxylic acid esters had a much greater confirmation rate than the acids.

The data of Figure 5 show the fallacy of using K value alone. Since K is unbounded (Figure 4), a molecule with many fragments will have a higher K value in matching than a molecule with fewer fragments. Thus, the carboxylic acids of carbon length >12 have a high K value even when the match is not good simply because of the number of fragments matched. One must also look at the negative points of the match.

In Figure 6 the relation between ΔK and confirmation can be seen. Note that again the use of retention data improves match confidence greatly at low match quality (high ΔK in this case). Note also that there is no anomalous behavior due to the carboxylic acids. In this case, since the difference between the calculated

match and a perfect match is represented, the effect of increased fragmentation due to increased molecular weight is not seen.

Observations such as this led to the use of K/K_{max} as the deciding match parameter. In this way, both positive and negative match parameters can be viewed in one term. In Figure 7, the relation of this parameter to match confirmation rate is shown to be a function of the size of K/K_{max}.

Conclusions. This study shows that the confidence of poor spectra matches can be greatly increased by using retention data as a match parameter. The confidence of excellent spectra matches is not affected by retention data -- probably because closely eluting compounds with very similar spectra begin to interfere at this level of confidence. The building of larger retention data libraries and the construction of an FT/IR segment to the historical library management program (Figure 8) are the next steps in improving the analyst's confidence in automated identifications.

REFERENCES

1. W. M. Shackelford, D. M. Cline, L. Burchfield, L. Faas, G. Kurth, and A. D. Sauter, Advances in the Identification and Analysis of Organic Pollutants in Water, L. H. Keith, ed., 527-554, Ann Arbor Science, 1981.

2. EPA-NIH Mass Spectral Data Base, John Wiley and Sons Registry of Mass Spectra, and ERL Athens Master Data Base.

3. G. M. Pesyna, R. Venkataraghavan, H. E. Dayringer, and F. W. McLafferty, Anal. Chem., 1976, 1362-1368.

4. EPA Contract Number 68-03-2867, Research Triangle Institute, Research Triangle Park, NC.

5. A. D. Sauter and D. Betowski, HRC&CC, 4, 1981, 366.

6. B. L. Atwater, Ph.D. Thesis, Cornell University, Ithaca, NY, 1980.

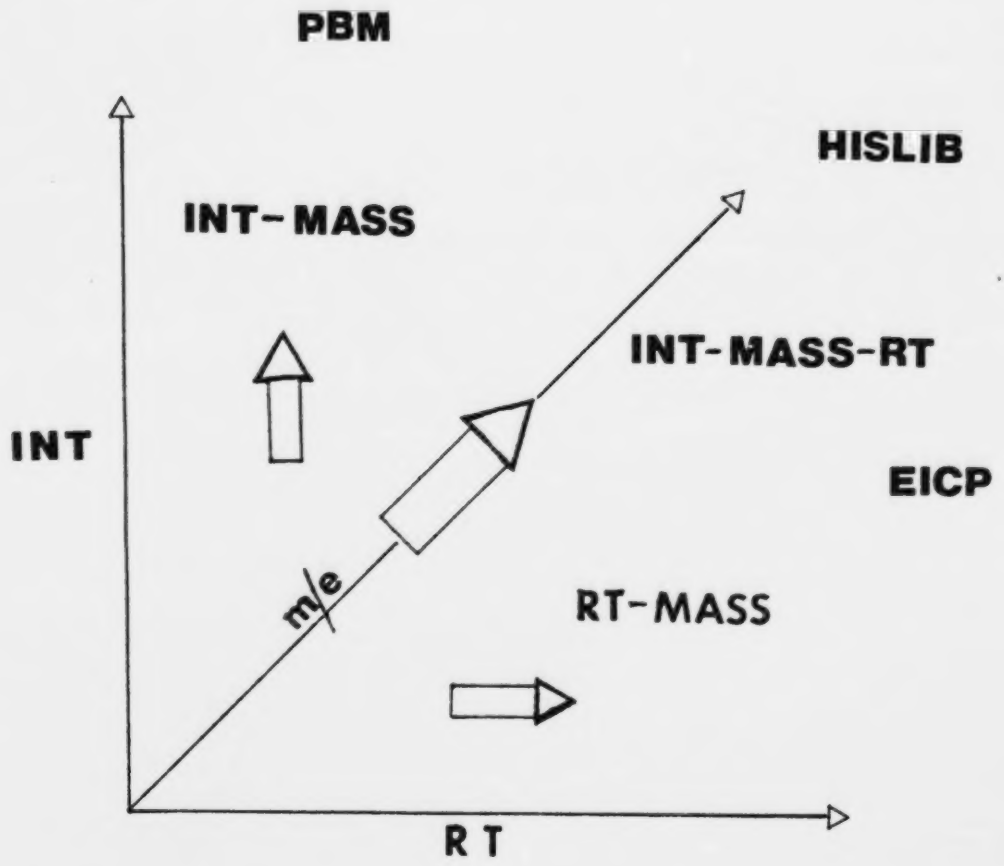

Figure 1. The three dimensions of GC/MS data.

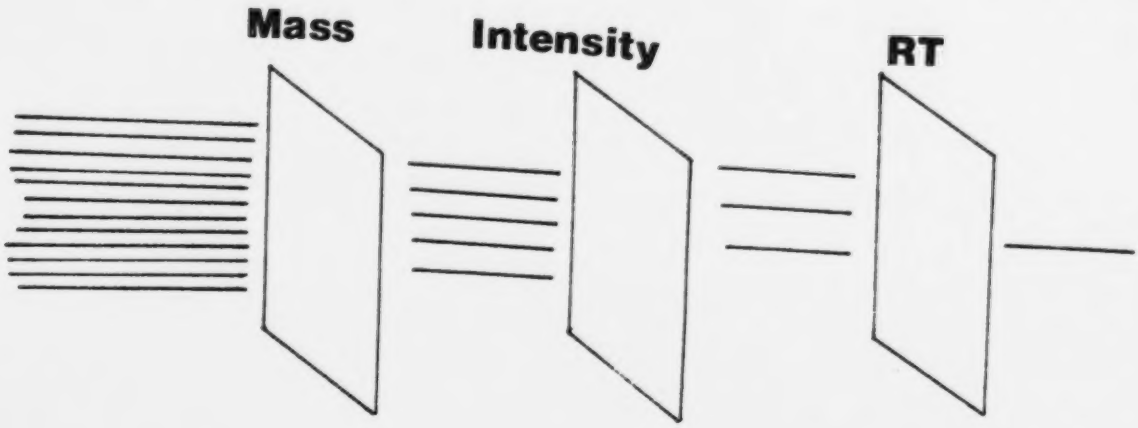

Figure 2. Mass, Intensity and Retention Time Filters for spectrum matching.

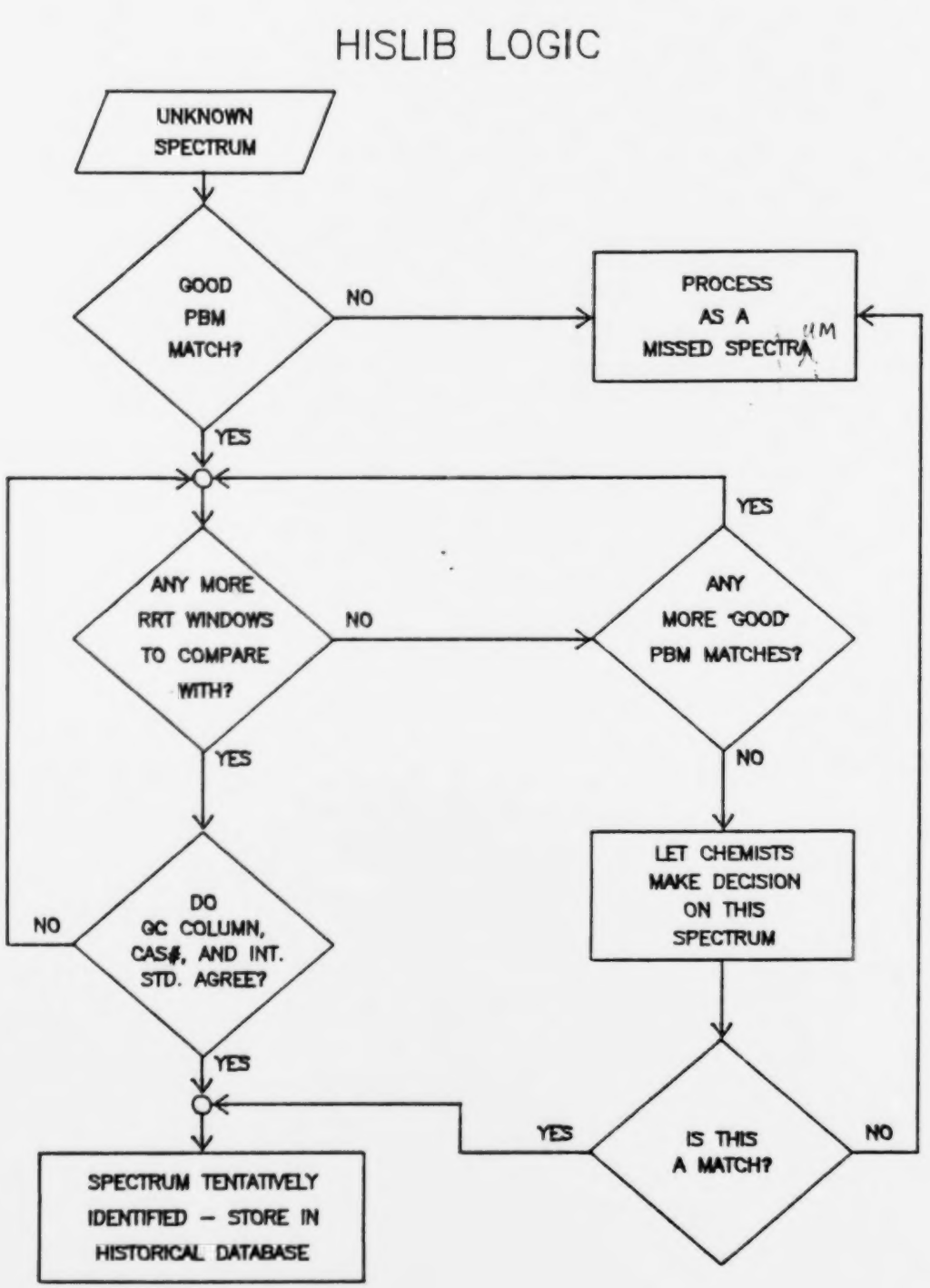

Figure 3. Logic of spectrum analysis in the historical library.

MATCHING SYSTEM PARAMETERS

PBM

- $K = \Sigma\ (U_i + A_i - D - W_i)$

  THEORETICALLY HAS NO LIMIT

- $\Delta K = K_{max} - K$

- $\frac{K}{K_{max}}$ MINIMIZES BIAS TOWARD COMPOUNDS WITH MANY FRAGMENTS

Figure 4. Probability based matching parameters. U_i = empirically derived uniqueness; A_i = empirically derived abundance value; D = dilution of spectrum by impurity; W = the tolerance allowed for abundance match.

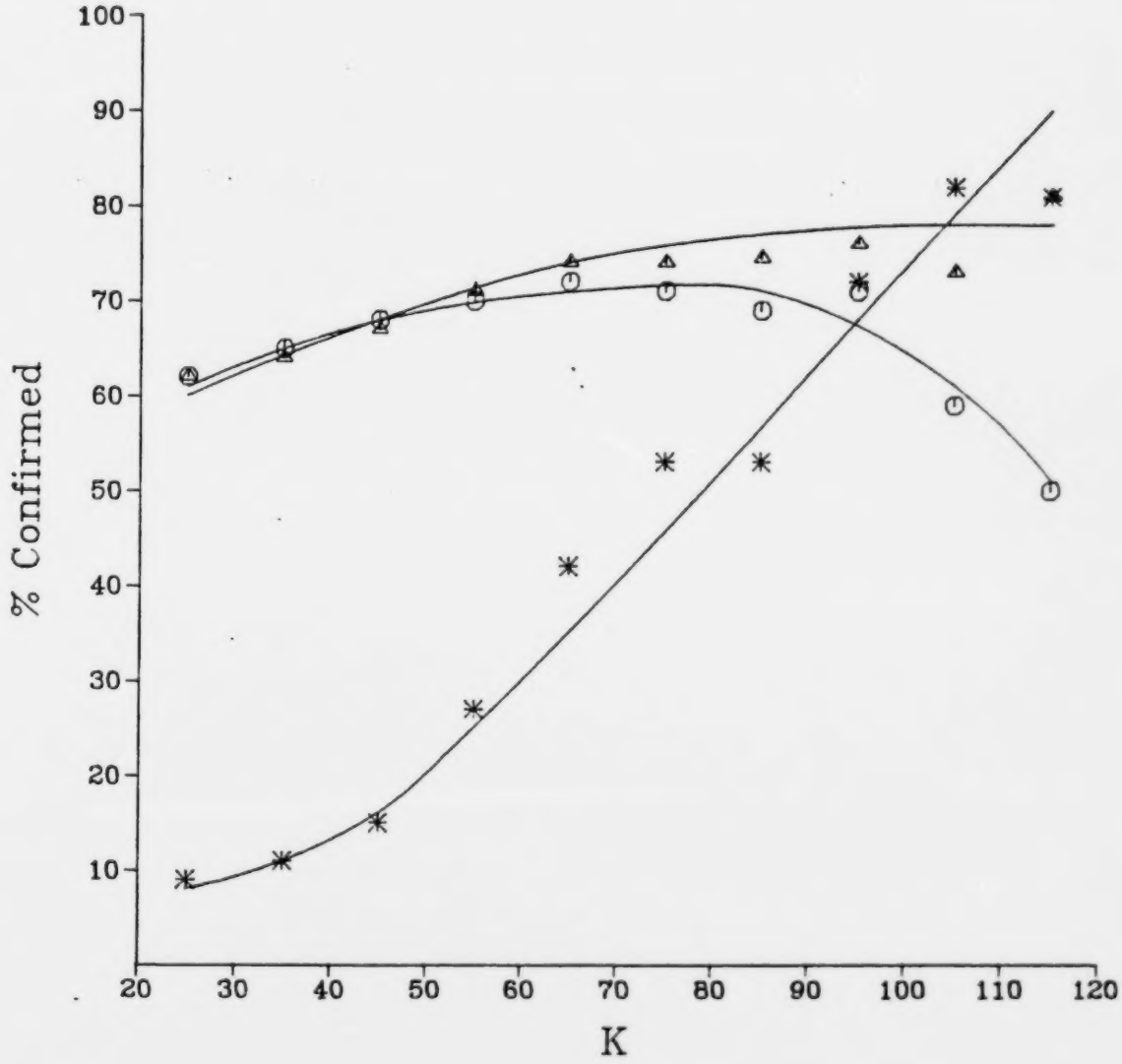

Figure 5. Plot of K versus confirmation rate for this study and that of Atwater (5).

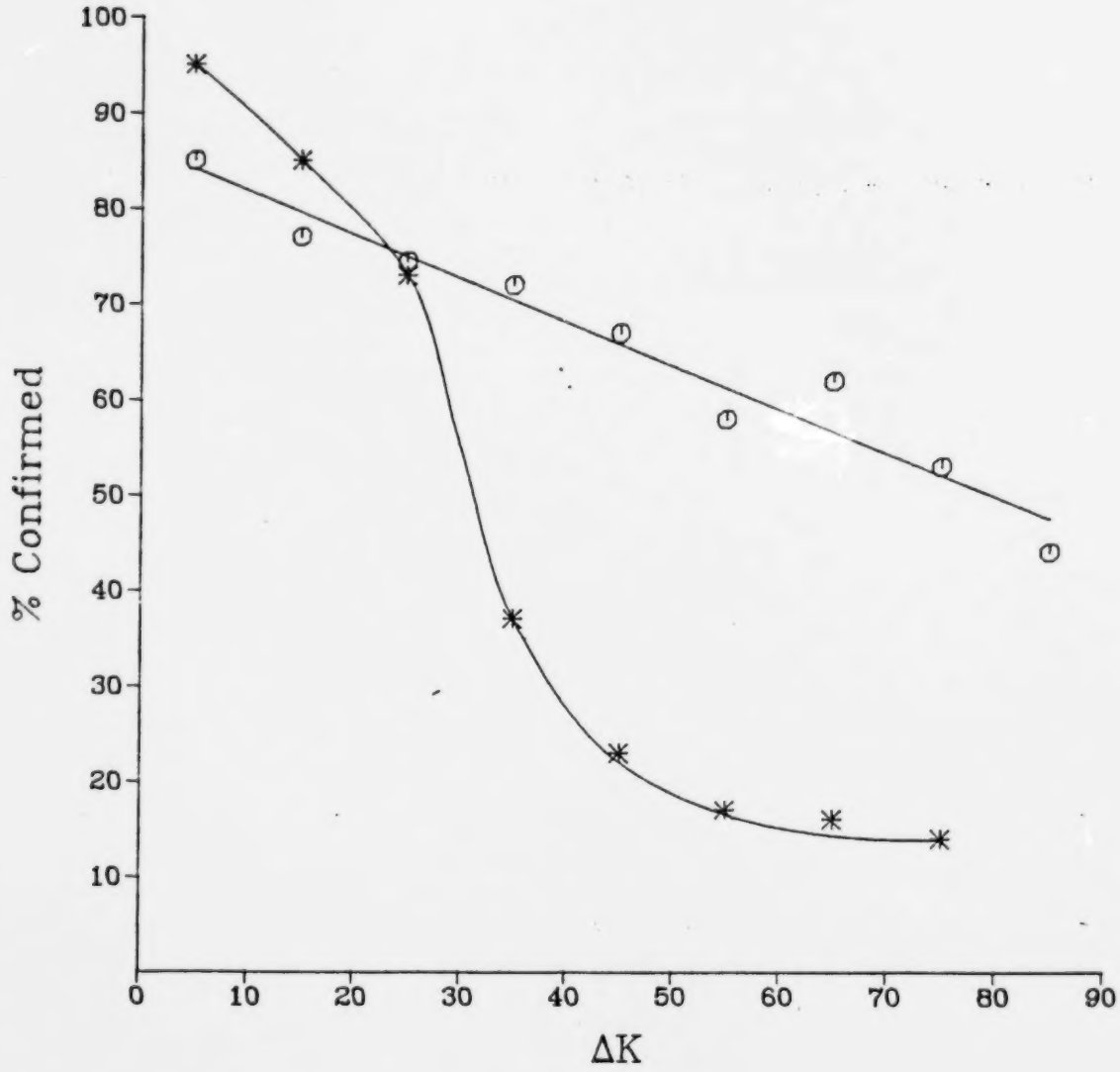

Figure 6. Plot of ΔK versus confirmation rate comparing this study with that of Atwater (5).

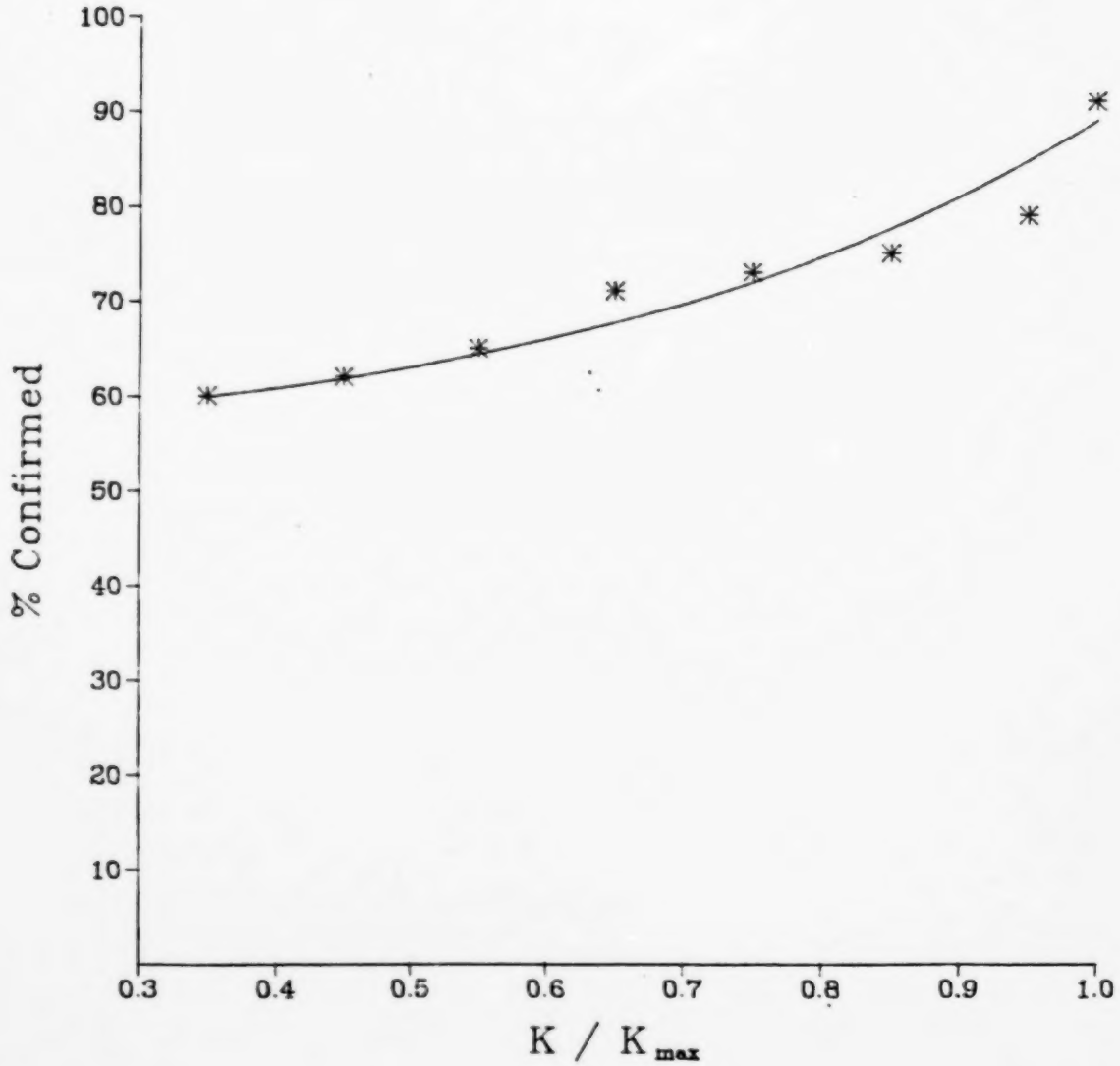

Figure 7. Relation of $\frac{K}{K_{max}}$ to confirmation rate with all compounds included.

FUTURE STUDIES

- BUILD RT LIBRARY WITH CAPILLARY COLUMN DATA
- ADD IR SPECTRA
- ADD SITE SPECIFIC DATA

Figure 8. Planned improvements to the historical library for further selectivity.

HOMOLOGOUS SERIES STANDARDS

C_6 - C_{19} N-ALKANES

C_{19} - C_{40} N-ALKANES

C_6 - C_{10} ALKENES

C_8 - C_{22} ALKENES

C_4 - C_{22} N-ALCOHOLS

C_3 - C_{16} ALDEHYDES

C_4 - C_{14} PRIMARY AMINES

C_4 - C_{18} SECONDARY AND TERTIARY AMINES

BENZENOID HYDROCARBONS

DICARBOXYLIC ACIDS

DIMETHYL ESTERS OF DICARBOXYLIC ACIDS

C_3 - C_{18} FATTY ACIDS

C_3 - C_{12} GLYCOLS

C_3 - C_{10} GLYCOL ETHERS

LOW BOILING ESTERS

C_3 - C_{19} METHYL KETONES

PHENOLS

PHTHALATE ESTERS

Table 1. Compound families used to initialize the historical library.

Table 2. Comparison of Retention Data with K and $\frac{K}{K_{max}}$ for Selected Compounds

Compound [CASRN]	Number of Occurrences	GC Column	Relative Retention Time Std. Dev.	K Median	K Range	$\frac{K}{K_{max}}$ Median	$\frac{K}{K_{max}}$ Range
2-butanone [78-99-3]	67	0.2% Carbowax 1500	± 0.062	36	62-20	0.57	0.86-0.41
1,1,2,2-tetrachloroethane [79-34-5]	93	0.2% Carbowax 1500	± 0.007	65	99-25	0.65	1.00-0.26
α-pinene [80-56-8]	82	0.2% Carbowax 1500	± 0.067	67	89-42	0.74	1.00-0.48
2-chlorophenol [95-57-8]	18	SP-1240 DA	± 0.043	52	78-38	0.56	0.79-0.41
benzoic acid [65-85-0]	298	SP-1240 DA	± 0.035	66	100-22	0.74	1.00-0.38
stearic acid [57-11-4]	66	SP-1240 DA	± 0.053	88	138-44	0.63	0.95-0.30
α-terpineol [98-55-5]	224	3% SP-2250	± 0.041	72	97-38	0.74	1.00-0.38
diethylphthalate [84-66-2]	222	3% SP-2250	± 0.012	75	113-36	0.66	1.00-0.32
cholesterol [57-88-6]	79	3% SP-2250	± 0.055	111	193-41	0.54	1.00-0.19
α-picoline [108-89-4]	21	SE-54 FSCC	± 0.007	59	74-43	0.64	0.80-0.46
nitrobenzene [98-95-3]	20	SE-54 FSCC	± 0.004	73	83-45	0.82	0.89-0.47
fluorenone [486-25-9]	10	SE-54 FSCC	± 0.001	87	95-55	0.92	1.00-0.58

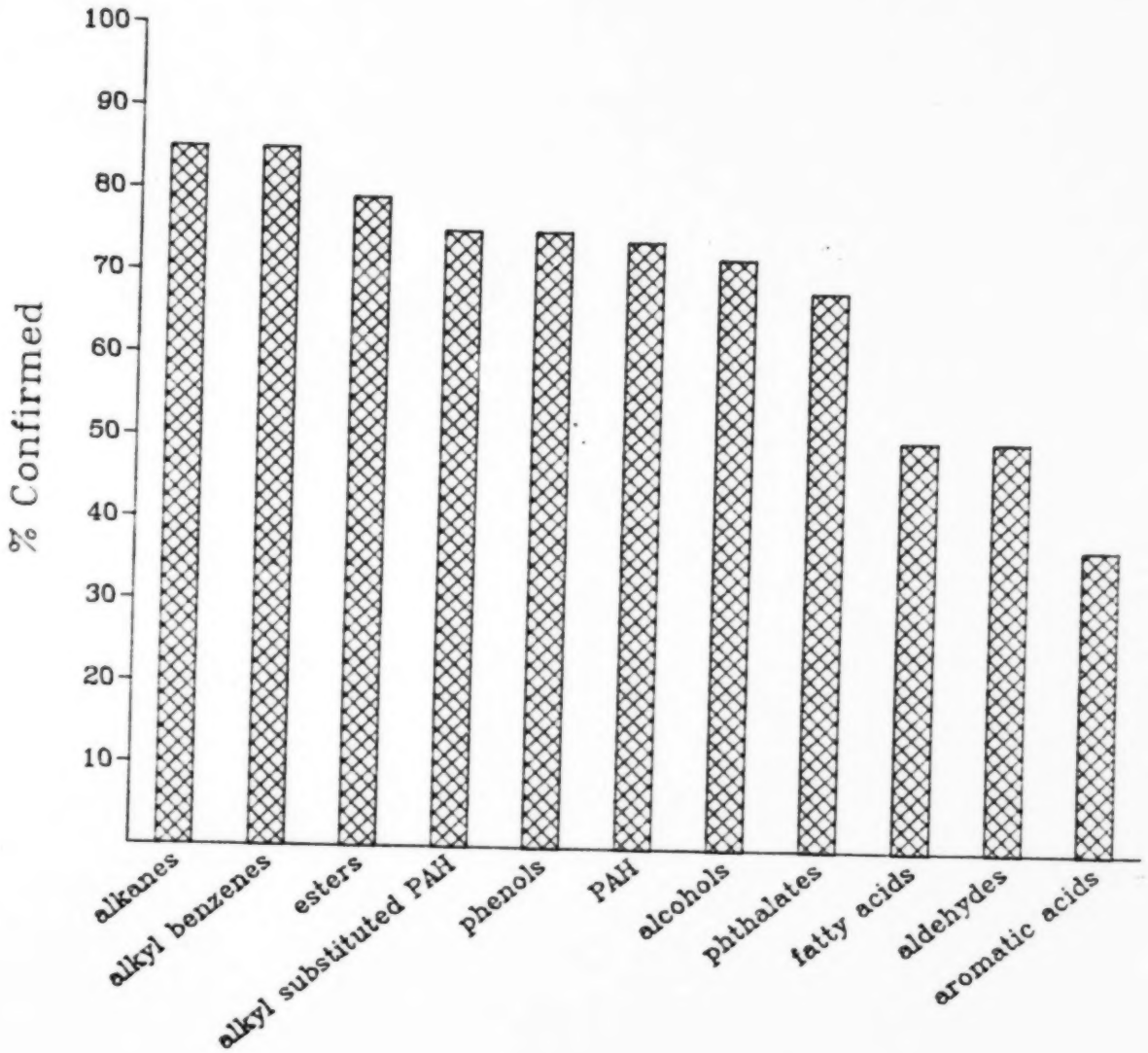

Table 3. Rate of confirmation for selected chemical classes.